Web 信息抽取技术在吹哨系统中的研究与应用

郑创伟　王　泳　陈少彬　邢谷涛　谢志成
（深圳市创意智慧港科技有限责任公司，广东　深圳　518034）

摘要：【目的】论述 Web 信息抽取技术在新闻舆情分析中的应用，为舆情虚假信息甄别、舆论引导提供新方法，从而避免对大众的思维、想法等造成不良影响。【方法】研究提出了基于行块分布函数和基于统计与网页结构两种不同的新闻正文信息抽取方法，使得在对 Web 新闻数据采集和存储的基础上，正文信息抽取更加高效和准确。【结果】两种 Web 信息抽取技术可以广泛应用于海量新闻数据分析、舆情监测等应用场景。【结论】通过基于行块分布函数的抽取方法和基于统计信息与网页结构的抽取方法，能够分别对轻量网页和大流量网页抽取信息时表现更优。

关键词：信息抽取；舆情；数据采集；分布函数；网页结构　　**中图分类号：**TP274　　**文献标识码：**A
文章编号：1671-0134（2023）04-154-05　　**DOI：**10.19483/j.cnki.11-4653/n.2023.04.032
本文著录格式：郑创伟，王泳，陈少彬，邢谷涛，谢志成．Web 信息抽取技术在吹哨系统中的研究与应用 [J]. 中国传媒科技，2023（04）：154-158.

导语

网络舆情是社会民意在互联网上的集中反映，其中 Web 新闻数据是当前数据采集的重要主体之一。针对网络上负面消极的舆论信息可能引发的舆情危机，媒体机构有责任建立一个舆情数据采集和分析机制。互联网数据所呈现的海量、多样、动态变化等特点，使得整个数据采集、管理与分析有着较大的困难，这也是当前政府和媒体亟待解决的重要问题之一。[1] 深圳报业集团创新研究的吹哨系统能够快速跟踪舆情变化趋势，从而全面了解舆情发展的来龙去脉。在吹哨系统中，主要运用了基于统计与网页结构的 Web 新闻正文抽取算法。因此，本研究针对多数据源网络舆情数据采集方法进行研究，对不同的 Web 新闻信息内容主题，设计了两类不同的新闻内容抽取算法，一类是基于行块分布函数的新闻正文抽取算法，另一类是基于统计与网页结构的新闻正文抽取算法。对两种方法进行对比，最终证明后者具有一定的优势，同时也反映出吹哨系统具有良好的可行性和一定的先进性。

1. 相关研究

1.1　Web 信息抽取技术分类

互联网数据绝大部分是以 HTML 文档的形式呈现，其文本信息、图片内容的位置都是无结构的，且这些数据都具有海量、异构数据源等特点。信息抽取是指针对非结构化的自然语言文本，利用相关技术从中筛选和提取有用的数据信息等，进而对其进行结构化转换，转换后可方便后续环节对其进行分析。但互联网信息数据有着极强的动态可变性和复杂性，内容更新频率非常高、数据量大，所需要的抽取技术较为复杂。[2]

为了解决以上信息抽取问题，当前国内外主要采取以下几种方式。[3]

1.1.1　基于规则模板

在网站布局时，虽然各网站的设计风格不同，但对同一个 Web 站点而言，往往会使用模板填充技术，使其网页具有相似结构。在抽取信息时，如果能够获取到该模板，就能快速、准确地取得相关数据源。但这种方式在后期维护时，要将不同的包装器定期加入模板库中，因此维护成本较高、可扩展性较差。

1.1.2　基于视觉特征

人们在日常浏览网页时，往往会根据一定的视觉特征来进行浏览。这就对网页内容和网页标签的布局等有了更高的要求。其中网页标签的作用较为重要，第一是用于组织网页的内容，第二是能够提供显示功能。因此在采集信息时，就可以根据这些具备一定视觉特征的分割页面进行抓取，从而提高采集效率。但随着互联网技术的发展，网页页面更加多样和丰富，导致提取这些视觉特征更为困难。

1.1.3　基于统计信息

在实际网页中，文本内容和网页标签等统计信息量不是均匀分布的，因此可以根据这些统计量的分布特征来制定不同的策略，进而实现 Web 信息抽取的目的，例如可将字符数作为正文区域的衡量指标等。这种方法有着较好的普适性，其最大的优势为不受数据源限制，而且在对数据抓取和学习时采用无监督学习方法，具有较好的效果。

1.1.4 基于 DOM 树结构的 Web 信息

在对某一网页信息进行抽取时，通过网页解析器可以将 Web 文档转化为 DOM 树，从而能够更加直接地看到 HTML 标签的层次结构。该种抽取方法目前也已较为成熟，主要抽取抓取后表格中大量带有节点特征的数据。这种方式综合考虑了网页整体结构和数据统计情况，因此在对新闻正文抽取时有着较好的表现。

1.2 Web 信息抽取结果衡量标准

在新闻信息抽取后，往往通过准确率 P 和召回率 R 来对评价结果进行衡量，公式如下：

$$P = \frac{系统正确抽取结果数}{系统所有抽取结果数} \times 100\%$$

$$R = \frac{系统正确抽取结果数}{系统应该抽取正确结果数} \times 100\%$$

从上述准确率 P 和召回率 R 的公式中能够看出，两者取值范围为 0–1，其数值越大则说明在信息抽取时更加准确。但需要注意的是，往往准确率 P 和召回率 R 两个指标不能同时增加或同时下降，一个指标的提升一般会使得另一个指标下降。[4]

2.Web 新闻数据采集与存储

在对 Web 新闻信息抽取前，需要对 Web 新闻数据进行采集，此时要考虑采集效率问题，保证可以快速采集到每天更新的海量 Web 新闻数据。

2.1 Web 新闻数据采集

分布式爬虫技术是当前较为成熟的一种数据采集方法，其分为对等分布式和主从分布式爬虫。第一种对等分布式爬虫在其运行时所有节点分工一致，参与爬虫任务的服务器都可以从待抓取 URL 队列中进行抓取，然后进行哈希处理，再对应到不同的爬虫节点上。但这种方法其拓展性一般，尤其在某个服务器出现问题重新分发任务时，会使得所有节点重新工作，造成了资源的浪费。第二种主从分布式爬虫则主要利用了一台服务器专门存储和处理 Master 节点，并将 URL 分发到 Slave 节点上，然后再进行后续数据采集工作。Master 节点不仅要完成分发任务，还需考虑分发后的 Slave 节点的负载均衡情况，这就对 Master 服务器性能提出了较高的要求，而这往往也是这种模式的瓶颈所在。[5]

本研究结合实际工作，采用主从分布式爬虫方式来采集数据，使用 Redis 数据库构建 URL 队列。将 Master 节点设定为控制节点，并利用 Redis 数据库管理 URL 爬虫队列，抓取后要存储 URL 链接以及对其进行去重处理。Slave 节点则从 Master 节点的 URL 仓库中获取网页链接，然后进行下载和解析，再将网页中的新闻内容存储在数据库。Spider 工作流程如图 1 所示。此外，Web 新闻网站每天都会更新大量新闻，为了能够保证采集到最新的新闻内容，在实际工作中还会定期注入人工种子 URL，主要包括需要采集的网站首页链接。

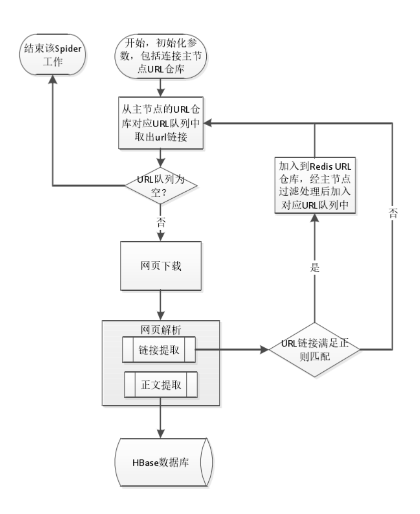


图 1 Spider 工作流程图

2.2 Web 新闻数据存储

新闻数据涉及的话题范围较为广泛且数据量较大，本研究建立数据存储中心，采用分布式数据库对其进行流式存储，构建数据存储中心主要包括以下内容。

2.2.1 构建数据存储平台

对于新闻数据的存储，主要是需要满足当前互联网技术爆发式的信息增长，考虑到信息内容的多样性，而传统的关系型数据库难以满足当下海量数据存储管理的实际需求。[6] 因此，本研究采用 Hadoop 分布式部署方式以及 HBase 分布式数据库来完成数据存储任务，并且利用虚拟化技术提升存储效率，在虚拟机中部署相关服务。通过这种方式，第一提升了非结构化数据存储能力，对于结构化数据将其转为二维关系表进行存储，提升其容错能力和查询效率。第二提升数据库并发效率，避免关系型数据库仅依靠事务机制来保证数据一致性这一限制，提升并发访问能力。

2.2.2 存储数据

使用 API 实现该部分功能，并且采集程序与数据存储相融合，支持海量数据高效装载到数据库中。为了保证分布式集群的负载均衡，将 Web 新闻网站的 URL 以“域名：协议：资源路径”的格式进行存储，然后完成数据抽取和结构化表示。数据库采用面向列的形式进行设计，表结构模式则可以依据列来进行确

认，可以根据实际需要添加字段信息或标签。在数据存储时，分布式数据库 HBase 底层存储使用 Hadoop 分布式部署方式，利用 ZooKeeper 实现多个任务的协同管理，并且能够利用分布批量计算能力来处理每天新闻产生的海量数据。采用主从分布式架构设计，其中 HDFS 作为底层存储实现，HMaster 则实现负载均衡。HBase 数据库是由 Table 和 Region 相互对应的，可以将其看作一张表，当数据量增加到一定程度时，Table 中的部分数据就会被分配到一个 Region 中。在实际存储过程中，HBase 的数据表就是以列的形式进行独立存储，并且会形成一个单独的文件，其中如果某一项值为空，那么该空值就会被舍弃，不会将其保存在数据库中。HBase 数据库适用于存储海量数据，这是因为其能够从横纵两个方向维度上支持数量级的弹性变化，一张单表就可以存储上亿条数据记录。另外，在数据检索时也支持列独立检索，当数据有一定关联时，可以将这些数据存储在同一列族下，这样就能够有效降低读写时 I/O 所消耗的资源，提升读写效率。[7] 因此，也就说明了这种列式存储方式能够更为方便地进行数据检索及更高效地压缩数据，从而更为适用非结构化数据的存储任务。

2.2.3 数据检索

在数据存储后，有时还需要对存储的数据进行检索查询，便于后续正文抽取与分析。在实际工作中发现，当数据量很大时，检索效率会受到影响，因此针对海量数据检索，本研究采用基于二级索引的方式来解决。二级索引主要是通过倒排的形式进行索引，对文本抽取的这一类新闻而言，主要是文本字符串数据信息，因此查询时主要依靠关键字信息，其原理如图 2 所示，例如对两列数据 C1、C2，C1 列用于描述 C2 列，C2 列为实际存储的图像数据，此时就可以在 C1 列建立二级索引，使用时可以直接查询 C1 列，然后提取到 C2 列数据信息的值。[8]

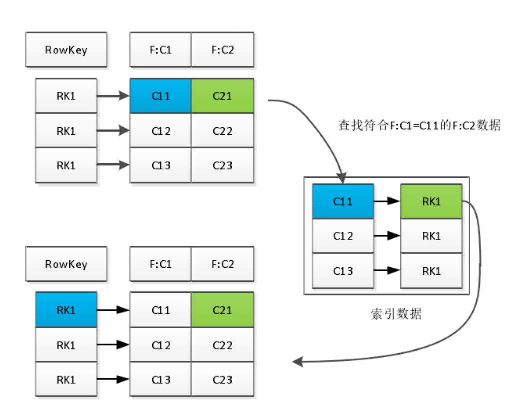


图 2 二级检索原理图

3.Web 新闻正文抽取方法研究

在对新闻网页进行抽取时，网页页面还包含了很多非正文噪声信息，例如导航栏、推荐广告等，那么在对正文内容抽取时，必须将这些信息进行过滤。而这些噪声信息往往包含在相同的网页结构中。Web 新闻网页中一般来说可将噪声信息分为两类，即可见噪声和不可见噪声。可见噪声主要包括导航栏、广告区等读者能够直接浏览的区域。不可见噪声主要指当打开网页源码时，会有 <script><style> 等标签信息，这些信息不会直接呈现给读者，但是当对文章进行抽取时就可能产生影响。

在创新研发的吹哨系统中，舆情分析的数据源主要来自 Web 网页中的新闻内容，主要包括新闻标题、发表时间、新闻正文内容三大要素。因此，以下分别介绍基于行块分布函数的 Web 新闻正文抽取算法和基于统计与网页结构的 Web 新闻正文抽取算法，本吹哨系统则最终选择后者，达到更佳效果。

3.1 基于行块分布函数的 Web 新闻正文抽取方法

基于行块分布函数的 Web 新闻正文抽取算法就是采用类似机器学习的思想，忽略了网页源码中相互交织的复杂和不规范问题，其核心要点主要是考虑网页正文区的文本密度和行块文本长度。其中，行块主要是指在去除源网页 HTML 标签后的空白行信息后，取本行上下各 2 行，共计 5 行为一个行块，用 RBT 表示。在抽取网页正文内容时，首先要建立行块分布函数，利用该函数，对各行块文本长度进行计算，寻找该区域内的骤升点和骤降点，并选出长度值较大的区域，则为网页的正文部分。通过该算法能够较好地抽取大多数 Web 信息网页信息，但当正文区域内容很短时，其受到噪声影响就更大，该算法有时就会出现误判情况，真正表述主题的正文内容就不会被抽取。针对这一点，主要是因为在正文内容中可能存在较多的逗号和句号，因此本研究考虑了中文标点所带来的差异，将行块文本长度限定于文本字符长度，可忽略 HTML 标签限制，在线性时间内通过行块分布函数来提取正文信息。[9] 算法公式如下：$RBTP_{RBi} = Len(c,p)$。

RBTP 即代表文本字符长度，RBi 代表编号为 i 的行块，c 代表该行块文本字符数，p 代表行块中的中文标点符号数。

网页正文抽取流程如图 3 所示。

3.2 基于统计与网页结构的 Web 新闻正文抽取方法

基于统计与网页结构的 Web 新闻正文抽取方法将充分考虑网页结构和 HTML 标签，Web 新闻网页包含

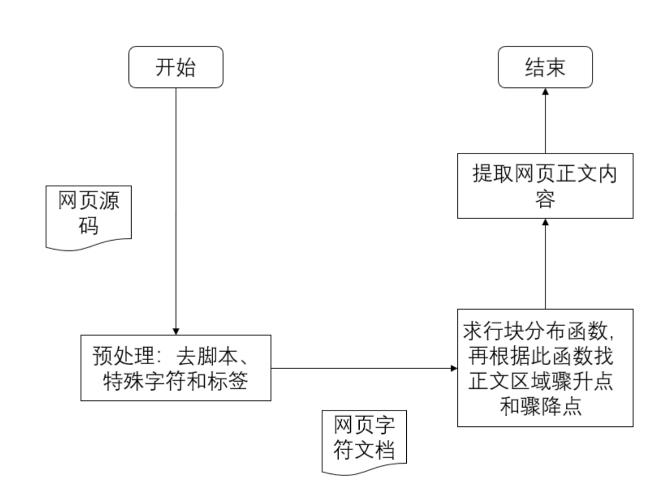


图 3　基于行块分布函数的网页正文抽取流程

了大量的文本信息，但有时噪声区域的文本长度较短，因此基于统计与网页结构的 Web 新闻正文抽取方法不再以文本长度来进行度量，其定义文本密度为标签元素节点内文本长度与该标签元素节点子树中标签元素节点的个数的比值：

$$ETD_{Ei} = \frac{C_{Ei}}{TN_{Ei}}$$

该公式可计算出标签元素节点的文本密度，属于递归算法的一种。代表不同标签元素节点的文本字符长度，则代表子树中标签个数。整个递归过程第一步是对网页信息进行预处理，主要是将源码中的标签信息剔除，例如 script、style 等，第二步则开始遍历 DOM 子数，统计标签元素节点信息。第三步则为计算文本密度，一般来说密度较高的区域就是正文区域。[10]

具体分步则如下：

（1）优化代码。考虑网页中整体布局和标签情况，对网页源码中用于布局和解释的标签相关代码予以删除，为后续遍历尽可能提供规律性较强的源代码；

（2）遍历子树。DOM 子树中含有大量标签元素节点信息，需要通过深度优先遍历的方式来采集数据信息。该方法从源码正文 <body> 区域开始，采取递归的方式逐一统计标签元素节点的信息，从而能够得到较为全面的新闻数据。

（3）计算文本密度。遍历子树后，要尽可能准确计算出每个标签元素节点的文本密度，Web 新闻正文区域则含有多个文本密度较大的内容块，如果文本密度较大，则说明该区域是正文区域。但在实际场景下，Web 新闻网页中含有较多的噪声，例如大量外链接等，此时就可能导致子节点内容块密度值大于父节点内容块密度值，如果此时对正文区域内容进行抽取，则可能导致抽取召回率过小。[11] 因此，为了找到包含 Web 新闻正文区域且深度最大的子树跟节点，引入标签元素节点文本密度和标签元素节点 EN 的文本密度公式如下：

$$ETDSum_{EN} = \sum_{i \in Childs_{EN}} EETD_i$$

其中，是元素节点 EN 的子节点集合，是其第 i 个孩子节点的扩展文本密度，这两个变量将其作为标签元素节点的两个属性。[12] 然后根据子节点集合情况以及子节点的扩展文本密度对节点进行打分：

$$Score_{EN} = \beta \times EETD_{EN} + (1-\beta) \times ETDSum_{EN}$$

其中，为权重因子，取值范围为 0–0.5，在通常情况下，正文区域下多内容块节点有较高的文本密度值。基于以上理论说明，该统计与网页结构的 Web 新闻正文抽取方法流程如下，流程图如图 4 所示。

（1）抽取算法开始，输入 Web 新闻文档；

（2）对文档信息开始进行预处理；

（3）执行递归操作，从网页主体 <body> 标签开始，逐个对元素节点进行遍历，再通过上述公式计算 ETDSum 值，即得到文本密度，最后对每个节点进行打分，得到 Scroe 值；

（4）再次从 <body> 标签进行递归遍历，找出 Scroe 值最大的节点，对其执行提取操作，提取出的文本内容则为新闻正文。

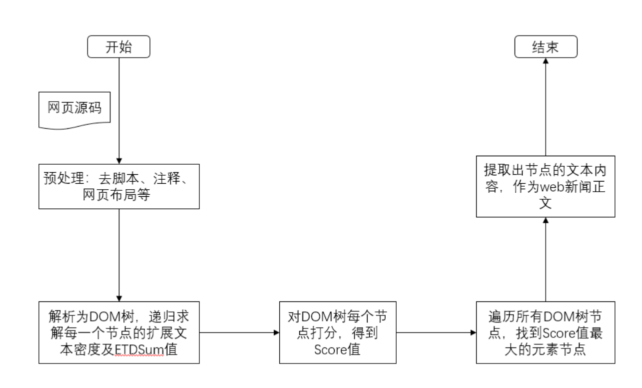


图 4　基于统计与 DOM 树的网页正文抽取流程

4. 实验结果及分析

以上抽取方法都属于一种无监督式的算法，下面从准确度和抽取速度两个指标来对比两种算法，分别从深圳新闻网、人民网、新浪新闻、新华网各爬取 200 个新闻网页。在这些 HTML 文档源码中，均含有 <title> 标签，内容主要为网页标题，其主要包括了新闻标题、出处以及时间，并显示在网页中的标题栏。在获取标题和时间时，本研究主要是提取这些信息后，

表 1 两种 Web 正文抽取算法抽取结果对比

新闻数据来源	基于行块分布函数的抽取方法		基于统计与网页结构的抽取方法	
	准确率 P	抽取速度	准确率 P	抽取速度
深圳新闻网	97%	2100ms	98%	1900ms
人民网	97.5%	2600ms	98.5%	2200ms
新浪新闻	96.5%	3600ms	99%	3800ms
新华网	97%	2200ms	99%	2100ms

利用正则表达式对字符进行分割和匹配。如“2019 年 03 月 13 日”“2021/04/18”之类的格式，都可以采用正则表达式进行匹配和抽取，表达式为：“（\d{4}[–，/，年]\d{1，2}[–，/， 月] \d{1，2}[日]？）（\s*）（\d{1，2}：\d{2}（：\d{2}）？）”。

在提取完标题数据信息后，主要需要对比正文抽取算法的优劣，其中抽取速度为多次测试后的平均值，准确率通过人工标注获得。信息抽取结果如表 1 所示。

根据以上结果可以看出，在信息抽取准确率上，基于统计信息与网页结构的抽取方法要优于基于行块分布函数的抽取方法。而在信息抽取速度方面，基于行块分布函数的抽取方法的时间复杂度为基于字符行的线性时间，在网页较小时抽取速度表现较好，但当网页较大时则基于统计与网页结构的抽取方法表现更优。所以，从网页抽取精度和速度两个方面综合考量，吹哨系统选择了基于统计和网页结构的 Web 新闻正文抽取方法。

结论

综上所述，在对 Web 新闻正文抽取时，基于统计信息与网页结构的抽取方法要优于基于行块分布函数的抽取方法，其准确率和抽取速度都更佳，非常适合当前网络数据传输量越来越大、网页信息越来越复杂的应用场景，最后通过实验测试，对两个方法进行了对比和分析。

参考文献

[1] 陈忱 . 大数据及融媒体技术在广电中运用论述 [J]. 中国传媒科技，2020（12）：78–80.

[2] 汤佳杰，曹永忠，顾浩 . 基于文本标点密度连续和的网页正文抽取 [J]. 计算机时代，2020（1）：69–72.

[3] 俞鑫，吴明晖 . 基于深度学习的 Web 信息抽取模型研究与应用 [J]. 计算机时代，2019（9）：30–32.

[4] 邱奇志，周三三，刘长发，陈晖 . 基于文体和词表的突发事件信息抽取研究 [J]. 中文信息学报，2018（9）：56–65，74.

[5] 魏春光 . 浅析互联网大数据在媒体业务的支撑应用——以人民公安报社舆情监测系统为例 [J]. 中国传媒科技，2019（6）：80–82，117.

[6] 王雪梅，陈兴蜀，王海舟，王文贤 . 基于标签和分块特征的新闻网页关键信息自动抽取 [J]. 山东大学学报（理学版），2019（3）：67–74.

[7] 陈俊洁 .Web 信息提取技术与应用的研究 [J]. 数字技术与应用，2017（9）：114，118.

[8] 袁然 . 全媒体传播中数据技术的应用实践 [J]. 中国传媒科技，2021（7）：21–23.

[9] 马晓慧，李泓莹 . 一种 DOM 树标签路径和行块密度结合的 Web 信息抽取方法 [J]. 智能计算机与应用，2017（4）：13–16，20.

[10] 胡露露，刘小勤，孙凯 . 基于正文特征和网页结构的网页正文抽取方法 [J]. 大气与环境光学学报，2017（3）：230–235.

[11] 王立志 . 网页信息抽取方法综述 [J]. 网络安全技术与应用，2022（3）：12–13.

[12] 赖娟，洪艳伟 . 基于规则约束的深度学习网络用于文本信息抽取 [J]. 计算机工程与设计，2021（12）：354–355.

作者简介：郑创伟（1978–），男，广东汕头，高级工程师，研究方向为大数据、人工智能；王泳（1977–），女，湖南邵阳，中级职称，研究方向为大数据；陈少彬（1973–），男，广东揭阳，中级职称，研究方向为大数据；邢谷涛（1984–），男，海南文昌，研究方向为云计算；谢志成（1980–），男，广东汕头，中级职称，研究方向为大数据、云计算。

（责任编辑：张晓婧）